

# Replicating Kernels with a Short Stride Allows Sparse Reconstructions with Fewer Independent Kernels

Peter F. Schultz,<sup>1,\*</sup> Dylan M. Paiton,<sup>2</sup> Wei Lu,<sup>3</sup> and Garrett T. Kenyon<sup>4</sup>

<sup>1</sup>*New Mexico Consortium*

<sup>2</sup>*University of California, Berkeley*

<sup>3</sup>*University of Michigan*

<sup>4</sup>*Los Alamos National Laboratory and New Mexico Consortium*

(Dated: June 18, 2014)

In sparse coding it is common to tile an image into nonoverlapping patches, and then use a dictionary to create a sparse representation of each tile independently. In this situation, the over-completeness of the dictionary is the number of dictionary elements divided by the patch size. In deconvolutional neural networks (DCNs), dictionaries learned on nonoverlapping tiles are replaced by a family of convolution kernels. Hence adjacent points in the feature maps (V1 layers) have receptive fields in the image that are translations of each other. The translational distance is determined by the dimensions of V1 in comparison to the dimensions of the image space. We refer to this translational distance as the stride.

We implement a type of DCN using a modified Locally Competitive Algorithm (LCA) to investigate the relationship between the number of kernels, the stride, the receptive field size, and the quality of reconstruction. We find, for example, that for 16x16-pixel receptive fields, using eight kernels and a stride of 2 leads to sparse reconstructions of comparable quality as using 512 kernels and a stride of 16 (the nonoverlapping case). We also find that for a given stride and number of kernels, the patch size does not significantly affect reconstruction quality. Instead, the learned convolution kernels have a natural support radius independent of the patch size.

## Introduction

Sparse coding has been widely used to model the structure of images. Typically a good sparse code requires learning an overcomplete dictionary of weights [2]. For image patches as small as 16-by-16 pixels, an overcomplete dictionary will still require hundreds to thousands of dictionary elements. It therefore would be advantageous to identify techniques that would reduce the number of independent weights that have to be learned.

For natural images, we expect image statistics to be similar at different parts of the image, and that important features of the image will be localized in space. This motivates the idea of deconvolutional networks [6], where the image is modeled as a sum of convolutions. Each kernel appearing in this sum captures a particular local image feature (for example, a Gabor filter), and is applied to a family of image patches, all of the same size but shifted relative to each other, by an amount we refer to as the stride.

Using deconvolutional networks can significantly reduce the number of free parameters used in learning the weights. This reduction occurs because the number of image patches that cover a given image pixel is greater than the number of kernels, due to the overlap in patches arising from the convolution. Hence fewer independent kernels are needed to achieve the same coverage of the image. The smaller the stride, the larger the effect of overlapping patches will be. In this work, we explore the relation between the stride, the number of convolution kernels, and the quality of the resulting image reconstructions.

## Background

In sparse coding, we seek to approximate an  $M \times N$  image  $\mathbf{x}$  in the form

$$\mathbf{x} \approx \sum_{k=1}^K y_k \phi_k, \quad (1)$$

---

\*Email address: pschultz@newmexicoconsortium.org

where  $\phi_1, \dots, \phi_k$  are  $M \times N$ -sized dictionary elements and  $y_1, \dots, y_k$  are scalars that constitute the representation of  $\mathbf{x}$ . By considering the  $y_k$  to be elements of a vector  $\mathbf{y}$  and the  $\phi_k$  to be column vectors of a  $(M \times N)$ -by- $K$  matrix  $\Phi$ , we can write

$$\mathbf{x} \approx \Phi \mathbf{y}. \quad (2)$$

In a sparse representation, only a small fraction of the  $y_k$  are nonzero. If  $K = MN$  there is typically a unique  $\mathbf{y}$  such that  $\Phi \mathbf{y} = \mathbf{x}$ ; however, this representation is unlikely to be sparse. When  $K = MN$ , close approximations are typically not sparse and sparse representations are typically not good approximations. If  $K < MN$ , the undercomplete case, it may be possible to find a close approximation, but it is even more unlikely that this approximation will be sparse. If  $K > MN$ , the overcomplete case, there are typically infinitely many solutions  $\mathbf{y}$  to the exact equation  $\Phi \mathbf{y} = \mathbf{x}$ , and we can expect that there are sparse representations  $\mathbf{y}$  that are good approximations to  $\mathbf{x}$ . The ratio  $K/MN$  is the overcompleteness factor. Typically, researchers have used dictionaries in the range of 0.75 to 10 times overcomplete [2], [3], [4], [5], [6], [7].

There are many ways to balance the closeness of the approximation with the sparseness of the representation. In this work, we will minimize an energy function

$$E(\mathbf{y}) = \frac{1}{2} \|\mathbf{x} - \Phi \mathbf{y}\|_2^2 + C(\mathbf{y}), \quad (3)$$

where  $C(\mathbf{y})$  is a cost function that penalizes nonsparse vectors  $\mathbf{y}$ . Here, we will use an approximation to the  $l^0$  norm using a threshold parameter  $\lambda$ :

$$C(\mathbf{y}) = \sum_k C_\lambda(y_k), \text{ with } C_\lambda(a) = \begin{cases} \lambda & \text{if } |a| \geq \lambda \\ 0 & \text{if } |a| < \lambda \end{cases} \quad (4)$$

Zeiler et al. introduced deconvolutional networks [6], in which we approximate  $\mathbf{x}$  by finding feature maps  $\mathbf{z}_j$  convolved with convolutional kernels  $f_j$ :

$$\mathbf{x} \approx \sum_{j=1}^J f_j * \mathbf{z}_j \quad (5)$$

The kernels  $f_j$  are small patches (for example, Gabor filters), and the  $\mathbf{z}_j$  are of size  $M \times N$ , with adjustments for edge effects.

The task is then to minimize the energy function

$$E(\mathbf{z}) = \frac{1}{2} \|\mathbf{x} - \sum_j f_j * \mathbf{z}_j\|_2^2 + C(\mathbf{z}). \quad (6)$$

It is natural to consider the collection of feature maps as an  $M$ -by- $N$ -by- $J$  grid above the  $M$ -by- $N$  image layer. Each gridpoint has a receptive field in the image layer  $\mathbf{x}$ , centered at the point below the voxel, and whose size is the same as the patch size of the  $f_j$ . Similarly, each pixel in the image influences a  $\text{size}(f_j)$ -by- $J$  region of the feature maps (figure 1). Thus, the feature maps play a similar role to the V1 layer in biological vision systems. Note that moving the blue point in the V1 layer by one pixel will move its receptive field by 1 pixel in  $\mathbf{x}$ -space; that is, the network has a stride of 1.

The  $f_j$  are analogous to the  $\phi$  in equation (3) and the  $\mathbf{z}_j$  are analogous to the  $\mathbf{y}$ . We note that deconvolutional networks can be described in terms of the formalism of approximation (1). Namely, there are  $K = JMN$  dictionary elements: for each of the  $MN$  pixels, translate each of the  $J$  convolutional kernels to be centered at that pixel. Accordingly, the overcompleteness factor is  $K/MN = J$ . Note that the overcompleteness does not depend on the patch size of the  $f_j$ .

Deconvolutional networks can also be extended to the case where the feature maps are downsampled from the original image size. In this case, the number of dictionary elements is determined by the number of pixels in the domain of  $\mathbf{z}_j$ . For example, if the  $\mathbf{z}_j$  are  $M/2$ -by- $N/2$ , the number of dictionary elements is  $J \cdot (M/2) \cdot (N/2)$  and the overcompleteness factor is  $J/4$ . Again, this factor does not depend on the patch size. We visualize the  $J$  feature maps as an  $(M/2)$ -by- $(N/2)$ -by- $J$  grid, but with a lower density of points than in the image layer, so that each gridpoint of the feature map lies above the center of its receptive field. Because of this lower density, moving one pixel in  $\mathbf{z}$ -space shifts the receptive field by 2 pixels in  $\mathbf{x}$ -space, giving a stride of 2. In general, feature maps with dimensions  $(M/F)$ -by- $(N/F)$  will have a stride of  $F$ .

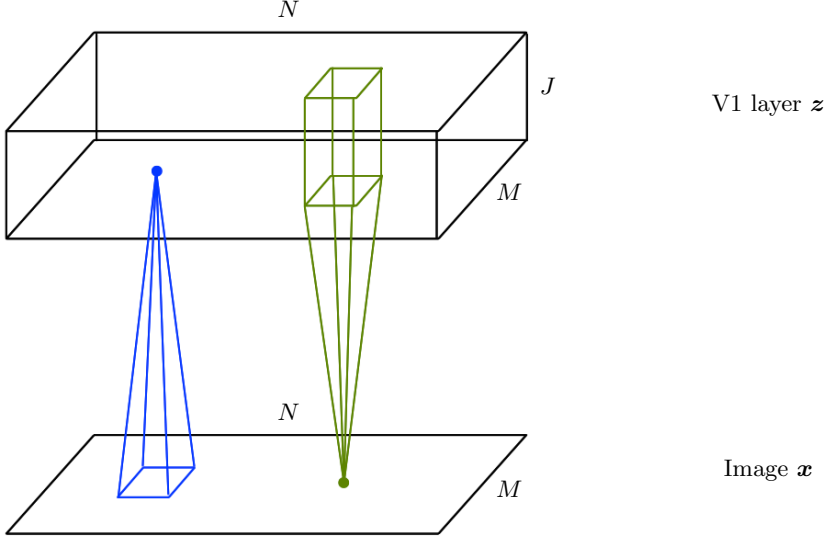


FIG. 1: The image layer  $\mathbf{x}$  and the V1 layer  $\mathbf{z}$ . Blue shows the receptive field in  $\mathbf{x}$  from a single point in  $\mathbf{x}$ . Green shows for a point in  $\mathbf{x}$  the region in  $\mathbf{z}$  whose receptive field contains the point.

For a deconvolutional network, each neuron is directly affected only by nearby neurons. Accordingly, we can use Locally Competitive Algorithms (LCA) of Rozell et al. [5] For a two-layer network consisting of an image  $\mathbf{x}$  and a V1-type layer  $\mathbf{y}$ , the LCA dynamics are as follows:

$$\frac{du_k}{dt} = -u_k + (\Phi^T \mathbf{x})_k - (\Phi^T \Phi \mathbf{y} - \mathbf{y})_k, \quad (7)$$

$$y_k = T(u_k). \quad (8)$$

Here  $u_k$  is an internal state variable corresponding to V1-neuron  $y_k$ , and  $T(u)$  is a transfer function. For the thresholded  $l^0$  cost function in equation (4),  $T(u)$  is the hard threshold function

$$T(u) = \begin{cases} u & \text{if } u \geq \lambda \\ 0 & \text{if } u < \lambda. \end{cases} \quad (9)$$

Equation (7) has an elegant motivation in terms of leaky integrators. The  $-u_k$  term provides leakiness of the internal state, the  $(\Phi^T \mathbf{x})_k$  term charges the neuron up based on input, and the  $\Phi^T \Phi \mathbf{y} - \mathbf{y}$  term provides local competition between different  $y$ -neurons. If the columns of  $\Phi$  have unit  $l^2$  norm, the  $-\mathbf{y}$  part eliminates self-interactions. This system will converge to a local minimum of the energy function (3). One advantage of LCAs is that the selection of local minimum is more stable as  $\mathbf{x}$  varies continuously, as in video. Also, the local nature of LCAs and DCNs means that the algorithm is well-suited to being implemented in hardware based, for example, on FPGAs or memristor arrays.

Note that in equation (7), we can group the two terms involving  $\Phi^T$ ; thus we see that the input  $\mathbf{x}$  enters the equation only as part of the residual,  $\mathbf{x} - \Phi \mathbf{y}$ , which is the part of the input that has not yet been accounted for by the representation  $\mathbf{y}$ . Our implementation therefore introduces an intermediate “residual layer” which holds the value  $\mathbf{r} = \mathbf{x} - \Phi \mathbf{y}$ . Our implementation is therefore given by

$$r_j = x_j - (\Phi \mathbf{y})_j \quad (10)$$

$$\frac{du_k}{dt} = -u_k + y_k + (\Phi^T \mathbf{r})_k \quad (11)$$

$$y_k = T(u_k). \quad (12)$$

sketched in figure 2. To improve computational speed, we truncate values in the residual layer to zero if their absolute value is below 0.005. In the context of convolutional networks with no downsampling, the term  $\Phi \mathbf{y}$  corresponds to  $\sum_j f_j * \mathbf{z}_j$ , and we consider  $\mathbf{z}$  to be an  $M$ -by- $N$ -by- $J$  layer of neurons. The transpose  $\Phi^T$  corresponds to convolutions with the reflections of  $f_j$ .

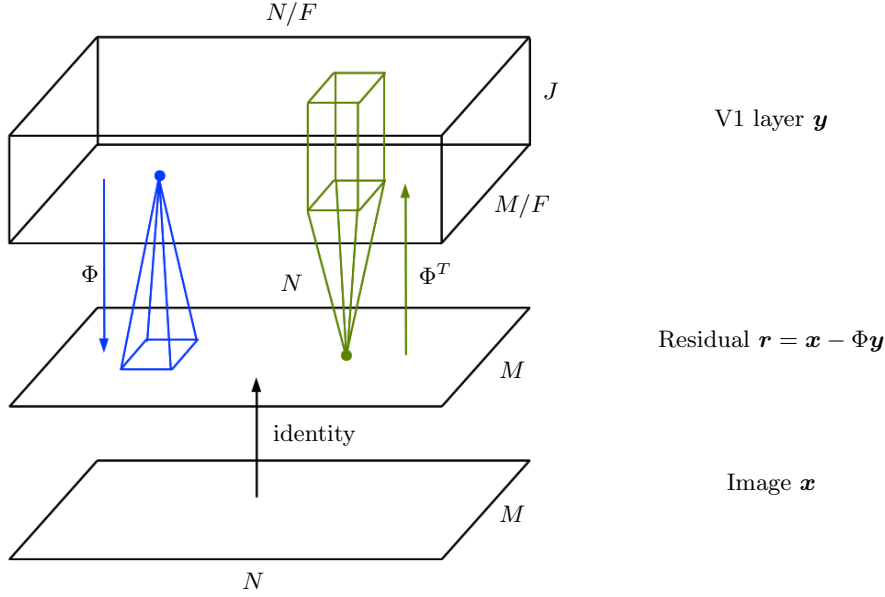


FIG. 2: Successful predictions by the V1 representation inhibits the residual layer by means of  $\Phi$ . Nonzero values in the residual layer excite the V1 representation by means of  $\Phi^T$ .

To learn the weights  $\Phi$ , we use stochastic gradient descent on equation (6). Writing the energy as

$$E = \frac{1}{2} \sum_p \left( x(p) - \sum_q f_j(q) z_j(p-q) \right)^2, \quad (13)$$

we see that the change in a kernel  $f_j$  at one point  $q$  is given by

$$\Delta f_j(q) \propto -\frac{\partial E}{\partial(f_j(q))} = \sum_p \left( x(p) - \sum_q f_j(q) z_j(p-q) \right) z_j(p-q) \quad (14)$$

$$= \sum_p r(p) z_j(p-q). \quad (15)$$

Note that the weights  $\Phi$  connect the feature maps  $z$  to the residual layer  $r$  (Figure 2). This is reminiscent of the Hebbian rule  $\Delta \Phi \propto r y^T$ . However, the fact that we are replicating patches across the feature maps means that the change in a weight is given by the sum over all replications of the kernel. Equation (15) has the form of a Hebbian rule. The sum over  $p$  reflects the fact that in the convolution, each weight is repeated over the image domain. The sum consists of each presynaptic/postsynaptic pair that is connected by the weight  $f_j(q)$ .

## Methods

We used PetaVision [1], an open-source neural network simulator that uses OpenMPI for parallel computation.

Our training set was taken from 482 Vine videos posted between Jan. 24 and Jan. 31, 2013. Each video was converted to a sequence of frames, for 79891 images, which were downsampled from 480x480 to 128x128 using GDAL. The images were then passed through a center-surround filter with mean 0 and  $\sigma$ -values of 0.5 for the center and 5.5 for the surround. For each choice of patch size, number of kernels and stride, we displayed each image for 200 timesteps, using a threshold of  $\lambda = 0.050$  and updating the weights at the end of the 200 timesteps. The initial value of the internal state  $u_k$  was random for the first frame, and for subsequent images, the initial state was the previous image's final state. Our training runs used a 64-process parallel computation on an AMD Opteron 6272-based machine.

At the end of the training run, we tested the dictionary by using it to generate sparse representations of the frames of one of the Vine videos, also downsampled to 128x128. Each image of the frame was shown for 200 timesteps. The entire video was repeated 10 times to eliminate startup artifacts. The representations obtained on the second pass were significantly different from that of the first, but the second through tenth passes were all substantially similar

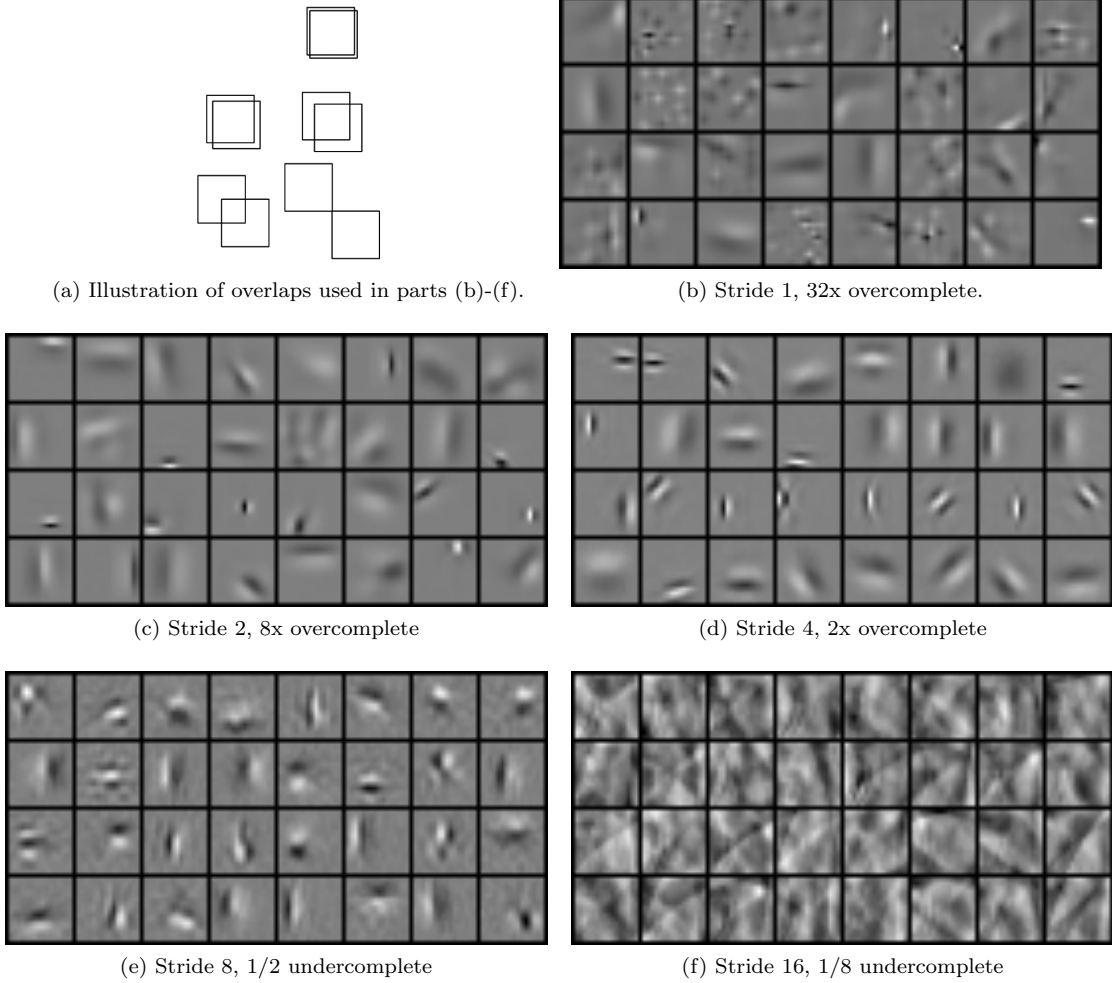


FIG. 3: Convolution kernels learned for varying strides with 32 convolution kernels.

to each other. Although all training runs were performed with a threshold  $\lambda = 0.050$ , reconstructions were run with threshold values of  $\lambda = 0.025, 0.050, 0.075, 0.100$ . The test runs were performed using a 16 processes on an AMD Opteron 8354-based machine.

We then evaluated the quality of the sparse representations by computing the fraction of V1 cells with zero activity (the percent inactive), and the  $l^2$  norm of the residual error, normalized by the  $l^2$  norm of the center-surround filtered image (the percent error). Good representations therefore have percent inactive close to 1.0, and percent error close to zero.

## Results

### Varying strides with fixed number of kernels

We trained several dictionaries using differing strides (1, 2, 4, 8 and 16) and 32 convolutional kernels. For each run, we used 16x16-pixel patch sizes, except for the stride 1 case, which used 15x15-sized patches. The reason for this difference is that a V1 neuron's receptive field should be centered on the V1 neuron. For odd strides, V1 neurons lie above image pixels, but for even strides, each V1 neuron lies above the center of a 2x2-pixel image patch. For stride  $F$ , the overcompleteness is  $32/F^2$ . Accordingly the problem should be overcomplete for strides 1, 2, and 4; and undercomplete for strides 8 and 16. In figure 3 we show the convolutional kernels learned in each of these runs. In the most overcomplete case, scale factor 1, there are several Gabor-like features, as well as several filters with high frequencies. This is consistent with a highly overcomplete dictionary. For scale factors 2 and 4, the resulting kernels

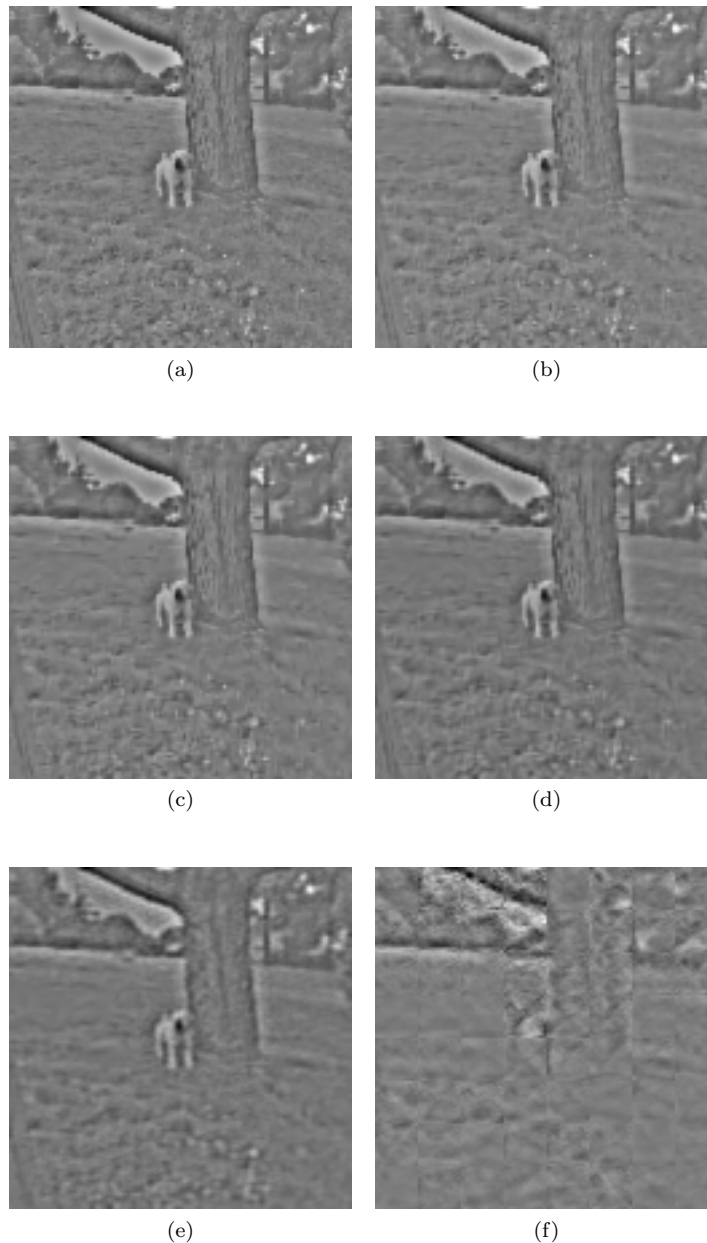


FIG. 4: Reconstructions from sparse representations using 32 kernels. Image (a) is the original image. Images (b)-(f) were generated using the dictionaries of the corresponding subfigure in figure 3.

are primarily Gabor-like. For a stride of 8, some Gabor-like features are visible but the quality of the kernels is visibly less than for the lower strides. Finally, for stride 16, the filters do not appear to show a useful dictionary.

Fig 4 shows reconstructions from the dictionaries. Part (a) shows the original image and (b)-(f) show the reconstructions using the kernels shown in the corresponding parts of figure 3. Parts (b)-(d), corresponding to overcomplete networks, show good reconstructions. Part (e) is slightly undercomplete; the image is recognizable but some loss in fine detail. Finally in part (f), the highly undercomplete case, the reconstruction is extremely poor.

In figure 5, we show plots of the error versus sparsity. Twenty runs are depicted: there are five strides (1, 2, 4, 8, and 16), and for each stride there are four thresholds (0.025, 0.050, 0.075, 0.100). For each run, the 196 frames of the video are shown as a point cloud, and the  $2\text{-}\sigma$  uncertainty ellipse is shown for that point cloud. Different scale factors are shown with different colors: blue for stride 1 runs, green for stride 2, red for stride 4, black for stride 8, and magenta for stride 16. Within a color, the lower thresholds have a lower percent inactive, and generally lower percent error. As the stride shrinks and the amount of overlap increases, we obtain greater and greater overcompleteness, and

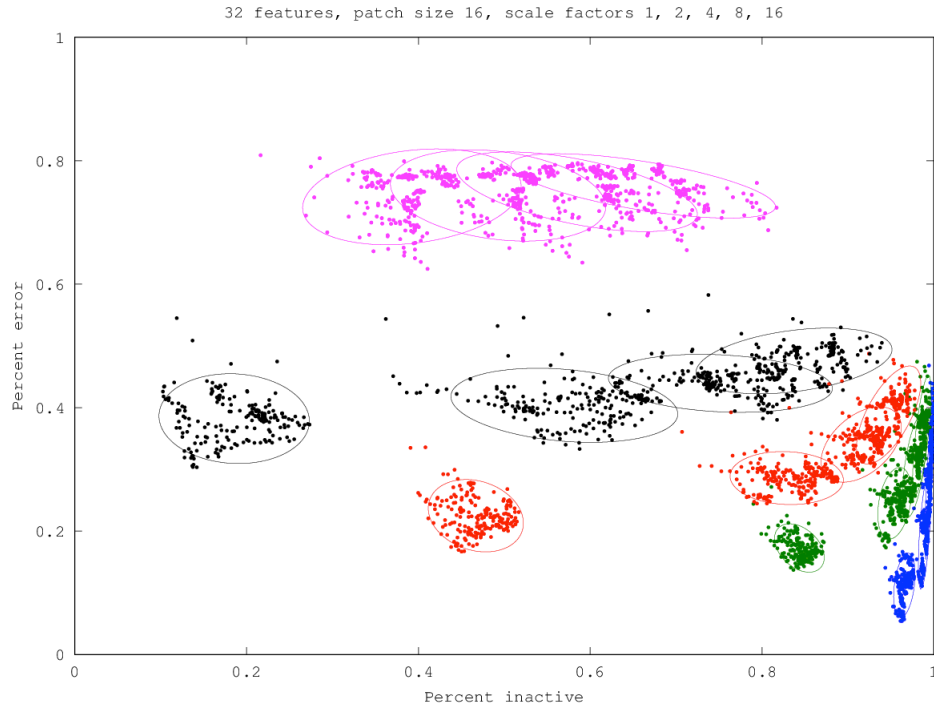


FIG. 5: Error versus sparsity plots for different strides, using a fixed number of features (32) and fixed patch size (16). Blue: stride 1, green: stride 2, red: stride 4, black: stride 8, magenta: stride 16. For each color, there are four runs, corresponding to  $\lambda = 0.025, 0.050, 0.075, 0.100$ . In each case the percent inactive increases as the threshold increases; generally error increases with increasing threshold.

hence sparser reconstructions with lower errors.

### Varying strides with fixed overcompleteness factor

For the next set of results, we used the same set of strides and the same patch sizes, and chose the number of kernels for each stride so that the dictionary was twice overcomplete. Thus, for stride 1, we used 2 kernels; for stride 2, 8 kernels, and so forth until for stride 16 (the nonoverlapping case), we need 512 kernels. In figure 6, we show the kernels learned. Figure 7 shows the reconstructions: they are all of approximately equal quality visually, although the higher strides preserve slightly more texture detail.

Figure 8 shows the plots of error versus sparsity. Although the stride-1 case (with only two feature maps) is slightly worse than the other strides, the ellipses for a given threshold show significant overlap. This result shows that by learning only 2-8 kernels and a small stride of 1 or 2, it is possible to approach the sparse reconstruction produced by 512 kernels that do not overlap.

### Varying patch size

We previously noted that for fixed stride and number of kernels, overcompleteness should be independent of patch size. In the final set of results, we test this prediction by measuring the effect of patch size on reconstruction quality. In figure 9, we show the kernels learned using  $8 \times 8$ ,  $16 \times 16$ ,  $32 \times 32$ , and  $64 \times 64$  patches. Although the patch sizes change, we see that the features being learned stay approximately the same size. Thus, the kernels learned in the  $64 \times 64$  experiment have a small region of nonzero weights; the larger patch size available does not lead to larger features being learned.

In Figure 10, we show the reconstructions from the kernels shown in figure 9. The visual quality of the reconstruction is similar across all four cases. This is confirmed in the plot of error versus sparsity in figure 11, where we see that for a given threshold, the four reconstructions using different patch sizes have substantially overlapping uncertainty ellipses.

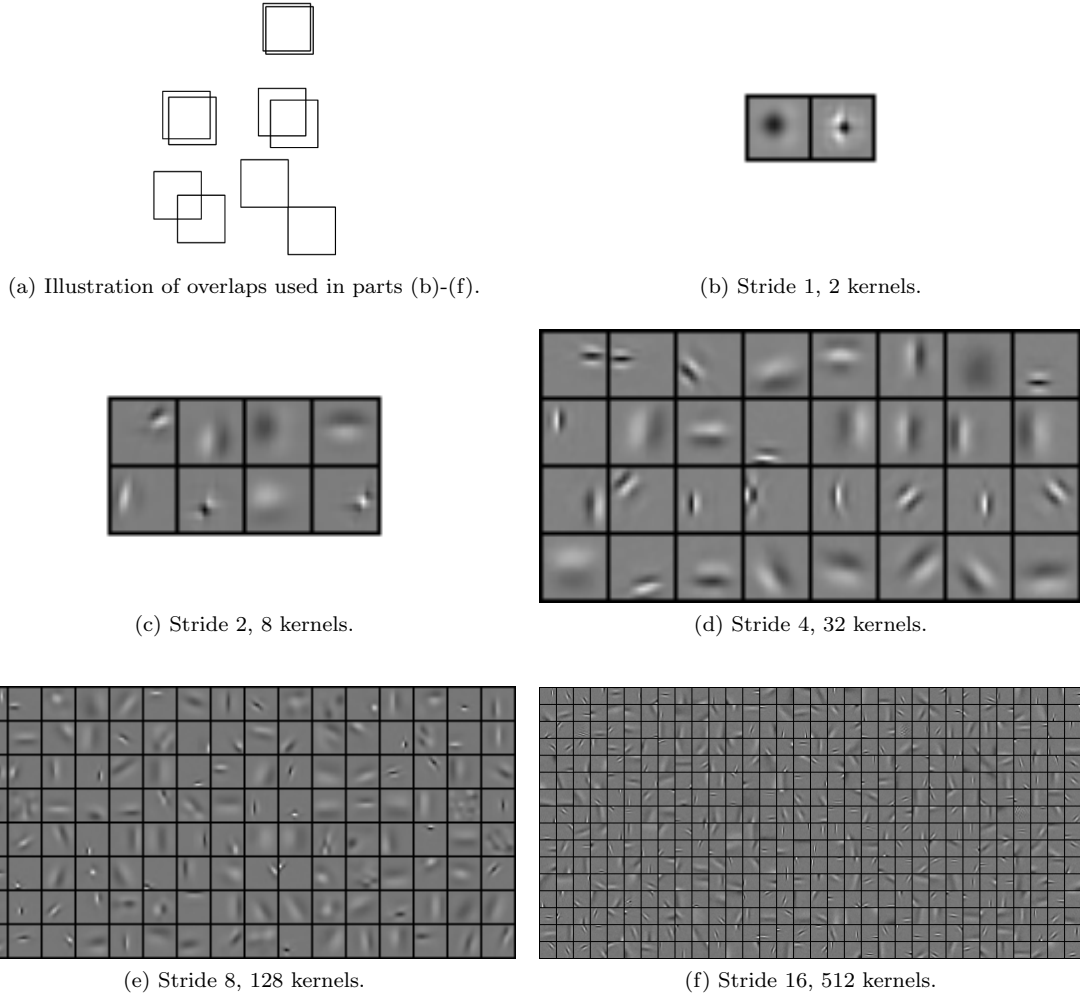


FIG. 6: Convolution kernels learned for varying strides with 2x overcompleteness. All patches are the same size, but the patches in (e) and (f) are shrunk to fit the figure on the page.

This confirms that over a wide range of patch sizes, the patch size has essentially no effect on overcompleteness or reconstruction quality, given a fixed stride and number of kernels.



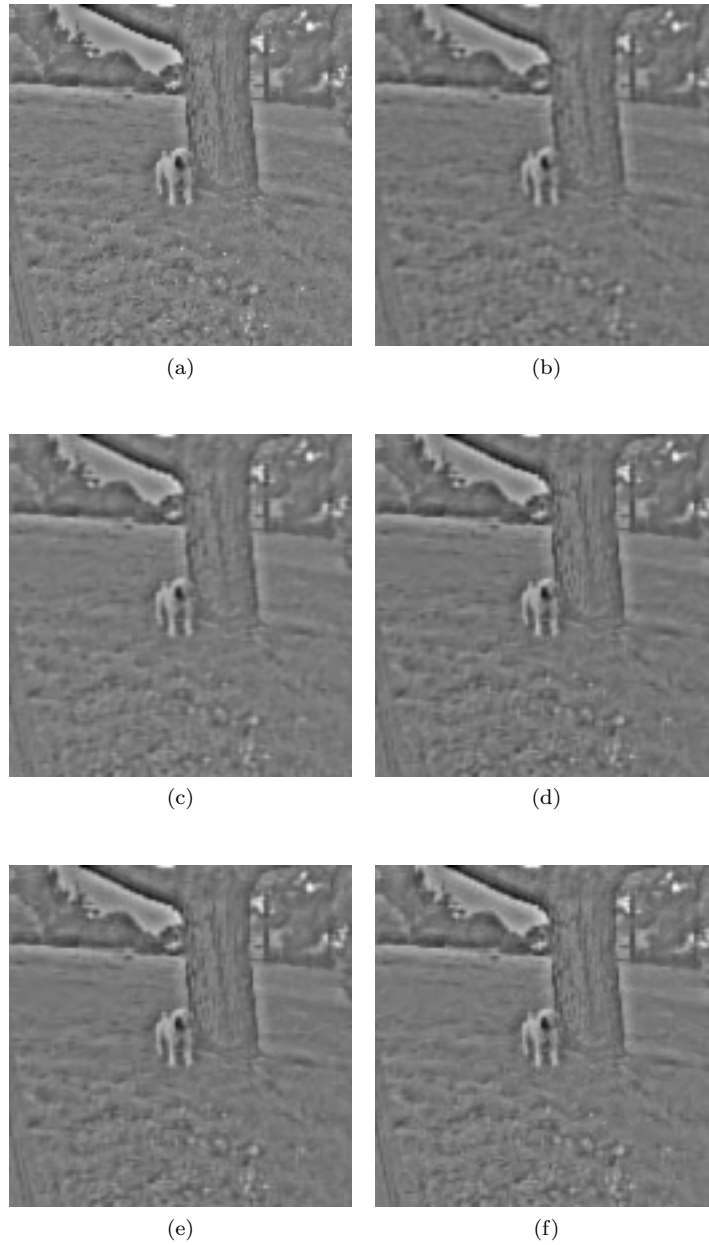


FIG. 7: Reconstructions from sparse representations using 2x overcompleteness. Image (a) is the original image. Images (b)-(f) were generated using the dictionaries of the corresponding subfigure in figure 6.

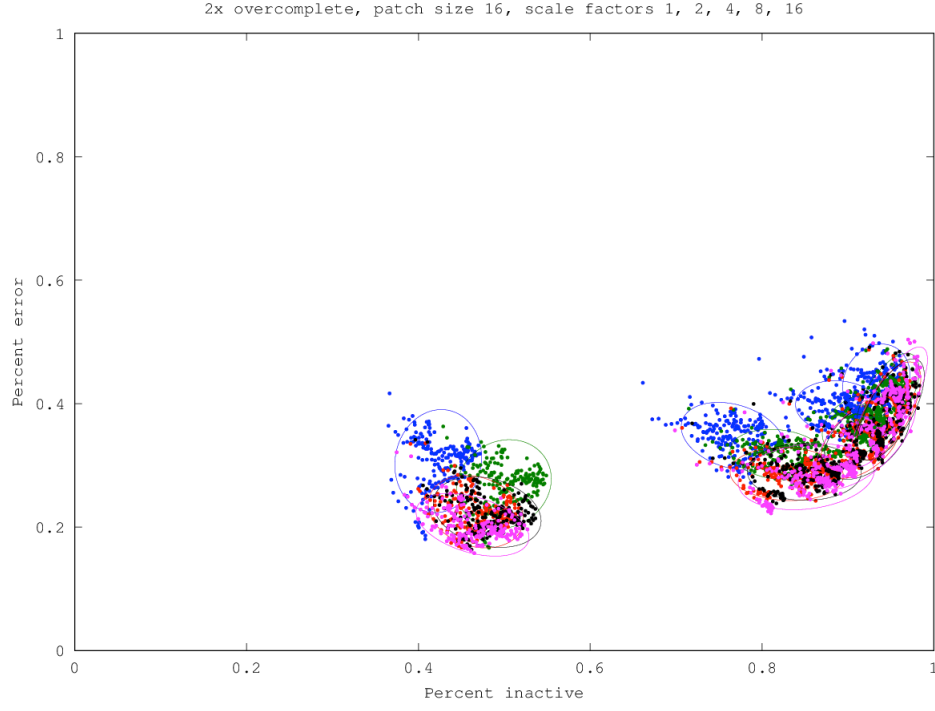
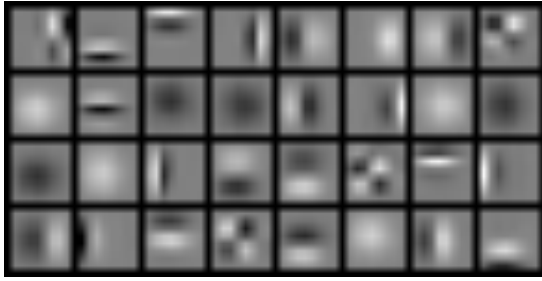
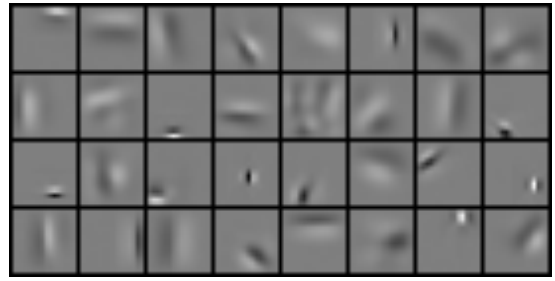


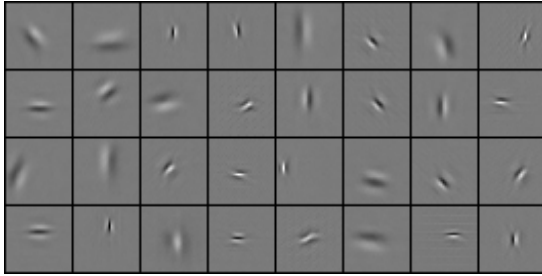
FIG. 8: Error versus sparsity plots for different strides, using a fixed overcompleteness (2x) and fixed patch size (16x16). Blue: stride 1, green: stride 2, red: stride 4, black: stride 8, magenta: stride 16. Values for threshold  $\lambda$  are as in figure 5.



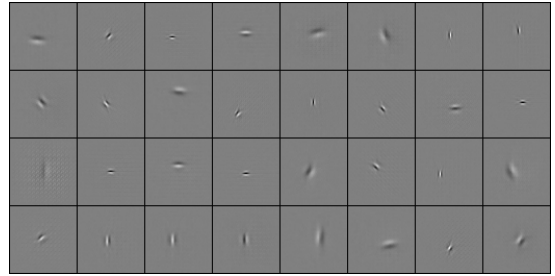
(a) 8x8 patches.



(b) 16x16 patches.



(c) 32x32 patches.



(d) 64x64 patches.

FIG. 9: Convolution kernels learned for varying patch sizes with stride 4 and 32 kernels.

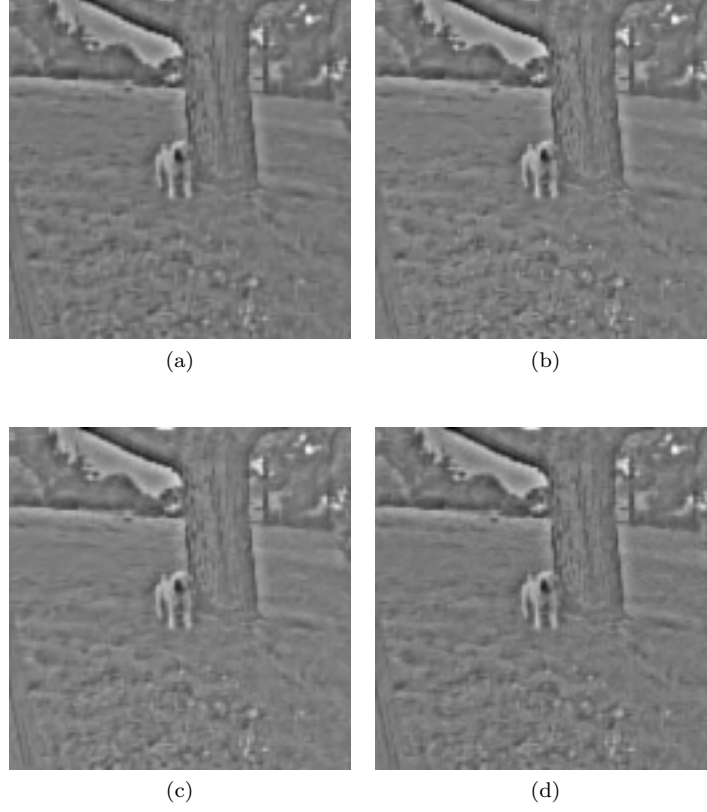


FIG. 10: Reconstructions from sparse representations. Images (A)-(D) were generated using the dictionaries of the corresponding subfigure in figure 9.

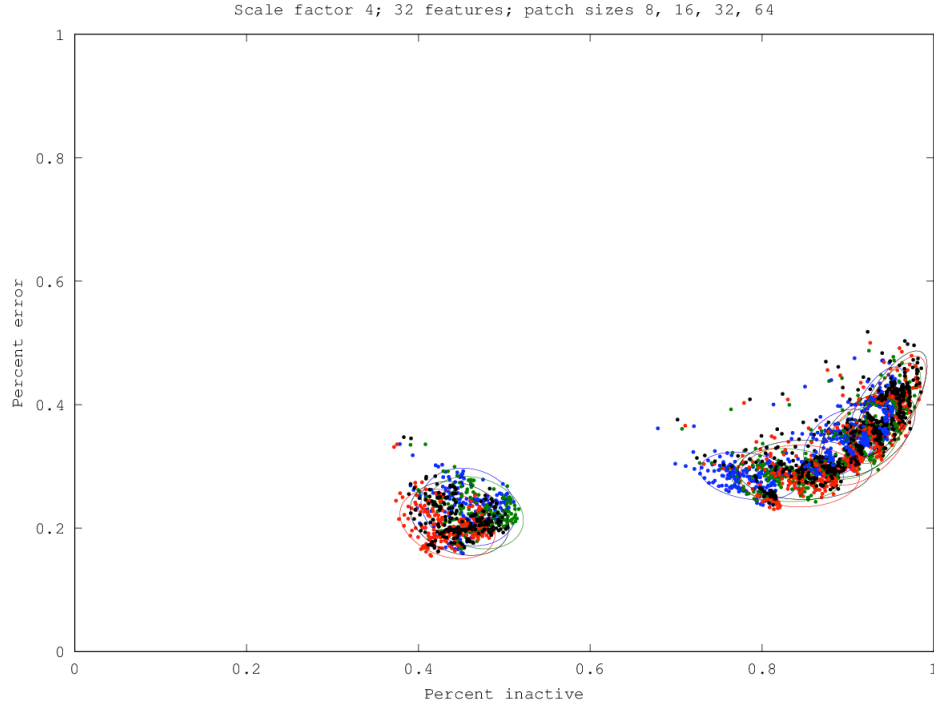


FIG. 11: Error versus sparsity plots for different patch sizes, using a fixed stride (4) and fixed number of kernels (32). Blue: 8x8, green: 16x16, red: 32x32, black: 64x64. Values for threshold  $\lambda$  are as in figure 5.

## Conclusion

We have seen that, for a dictionary learned from a deconvolutional network, the overcompleteness and quality of reconstruction is determined by the stride and the number of features, and not the patch size. Indeed, even for large patches, the learned filters tend to have small regions of support. Since the number of independent parameters in the dictionary is given by the patch size and the number of features, we observe that we can increase overcompleteness (and hence the quality of the sparse representation) without increasing the number of parameters by increasing the overlap of the receptive fields of adjacent V1 neurons.

## Acknowledgments

Work performed for the DARPA UPSIDE Program under Cooperative Agreement Award HR0011-13-2-0015.

- 
- [1] Petavision. <http://sourceforge.net/p/petavision/code/HEAD/tree/>.
  - [2] Bruno A. Olshausen. Highly overcomplete sparse coding. In *Proc. SPIE*, volume 8651, pages 86510S–86510S–9, 2013.
  - [3] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 681:607–609, 1996.
  - [4] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.
  - [5] Christopher J. Rozell, Don H. Johnson, Richard G. Baraniuk, and Bruno A. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20:2526–2563, 2008.
  - [6] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535, 2010.
  - [7] Joel Zylberberg, Jason Timothy Murphy, and Michael Robert DeWeese. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Computational Biology*, 7(10), 2011.